علوم دامی ایران
دورهٔ ۴۷، شمارهٔ ۲، تابستان ۱۳۹۵ (ص ۲۹۰-۲۷۹) DOI: 10.22059/ijas.2016.59033

گروه‌بندی ژنتیکی گاومیش‌های بومی آذری و شمالی با روش شبکهٔ عصبی SVM

زهرا عزیزی[۱]، حسین مرادی شهربابک[۲*]، محمد مرادی شهربابک[۳]، عباس رأفت[۴] و جلیل شجاع[۵]
۱، ٤ و ۵. دانشجوی دکتری، دانشیار و استاد، گروه علوم دامی، دانشکدهٔ کشاورزی، دانشگاه تبریز
۲ و ۳. استادیار و استاد، گروه علوم دامی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج
(تاریخ دریافت: ۱۳۹٤/۱۰/۱۵- تاریخ تصویب: ۱۳۹۵/۱/۲۵)

چکیده

هدف این تحقیق گروه‌بندی گاومیش‌های استان‌های آذربایجان شرقی، غربی و اردبیل از بوم‌جور (اکوتیپ) آذری و استان گیلان از بوم‌جور شمالی و درنهایت قابلیت جداسازی افراد مناطق مختلف با روش یادگیری ماشین بود. به شمار ۲۵۸ گاومیش از مناطق مختلف دو بوم‌جور شمالی و آذری نمونه‌گیری شد و با استفاده از SNPChip 90K مربوط به شرکت افی متریکس در کشور ایتالیا تعیین ژنوتیپ شد. برای پیش‌بینی عملکرد روش ماشین بردار پشتیبان برای گروه‌بندی افراد، دو روش متریک اعتبارسنجی متقابل و سطح زیر منحنی مشخصهٔ عملکرد سامانهٔ (AUC) اعمال شد. نتایج آزمون اعتبارسنجی متقابل و سطح زیر منحنی برای گروه‌بندی افراد چهار منطقه به ترتیب ۹۲ و ۹٦ درصد بود که گویای این است که باوجود نزدیک بودن افراد گله‌های مختلف و سخت بودن جداسازی این افراد، روش ماشین بردار پشتیبان با درستی بالایی، توانایی اختصاص دادن افراد به گله‌های مربوط به خود را دارد. نتایج آزمون‌های اعتبارسنجی متقابل و سطح زیر منحنی مشخصه عملکرد سامانه برای دو بوم‌جور به ترتیب برابر ۹٦ و ۹۸ درصد بود که نشان‌دهندهٔ قابلیت جداسازی بهتر دو بوم‌جور است. روش یادگیری ماشین با توجه به این موارد و با پیش‌بینی‌هایی که برای گروه‌بندی هر فرد انجام می‌دهد می‌تواند در کنترل کیفیت و کاربردهای ژنتیکی کارآمد باشد.

واژه‌های کلیدی: SNPChip 90K، گاومیش، گروه‌بندی، ماشین بردار پشتیبان.

مقدمه

گاومیش یک‌گونه از حیوانات اهلی بسیار باارزش و چندمنظوره است که در برخی از نواحی ایران پرورش داده می‌شود. در ایران بر پایهٔ آمارنامهٔ مرکز آمار ایران حدود ۱۹۹۰۰۰ گاومیش وجود دارد که بیشتر آن‌ها در جنوب و شمال غرب ایران نگهداری می‌شوند. همهٔ گاومیش‌های ایرانی، در دستهٔ گاومیش‌های رودخانه‌ای قرار دارند (Anonymous, 2012). گاومیش‌های ایران بر پایهٔ شرایط آب و هوایی در سه دستهٔ اصلی بوم‌جور (اکوتیپ) آذری (آذربایجان غربی و شرقی)، ۲) بوم‌جور شمالی (گیلان و مازندران) و ۳) بوم‌جور خوزستانی (خوزستان) تقسیم می‌شوند. یکی از چالش‌هایی که همواره درست بودن بررسی‌های GWAS[1] را تهدید می‌کند، چالش ساختارهای زیرجمعیتی درون جمعیت‌های موردبررسی است. به‌عنوان‌مثال، اگر در یک پژوهش Case-Control یکی از دو جمعیت موردبررسی، متعلق به یک نژاد و جمعیت دیگر از نژادی دیگر باشد، همواره این امکان وجود دارد که جایگاه‌های ژنتیکی معرفی‌شده با اثر معنی‌دار، منشأ گرفته از تفاوت در صفت موردبررسی نباشند، بلکه

* تلفن: ۰۲۶۳۲۲۴۸۰۸۲
E-mail: hmoradis@ut.ac.ir
1. Genome Wide Association Study

ناشی از تفاوت‌های نژادی دو جمعیت یادشده باشند. لذا باید سعی شود که جمعیت‌های انتخاب‌شده در بررسی‌ها تا حد امکان برای دیگر صفات همگن باشند و تفاوت‌های آن‌ها تنها در صفت موردبررسی باشد (Thomas & Witte 2002; Wacholder *et al*., 2002). از سوی دیگر نشانگرهای ژنتیکی می‌توانند برای شناسایی و تأیید منشأ افراد استفاده شوند. ریزماهواره‌ها باوجود برتری‌هایی مانند چندریختی زیاد، داشتن پراکنش مناسب در سراسر ژنگان (ژنوم)، به دلیل اشتباه در تعیین اندازه و فراخوانی اندازهٔ همردیف ژنی (اللی)، نتایج به‌دست‌آمده از این نشانگرها مناسب نبوده و تجزیهٔ ریزماهواره‌ها حتی باوجود نرم‌افزارهای موجود زمانبر است (Fernández *et al*., 2013). بین دامنهٔ گسترده‌ای از نشانگرهای مولکولی که توسعه یافته‌اند، چندریختی‌های تک نوکلئوتیدی (SNP)ها[1] بیشترین فراوانی را دارند و در سراسر ژنگان به‌طور گسترده‌ای پراکنده‌اند و توزیع این SNPها بین گونه‌ها متفاوت است (Vignal *et al*., 2002).SNP ها در بررسی تنوع ژنتیکی دام‌های اهلی و ساختار جمعیت سودمند هستند (Lin *et al*., 2010). نتیجه‌گیری ساختار جمعیت از نشانگرهای ژنتیکی، در شرایط گوناگون مانند بررسی‌های ارتباطی و تکاملی (Bridges *et al*., 2011)، دسته‌بندی زیرگونه‌ها و تعیین بازدارنده‌های ژنتیکی (نواحی جغرافیایی که با تغییر ژنتیکی شایان‌ملاحظه‌ای ارتباط دارد (Guerard *et al*., 2004)) سودمند است (Lao *et al*., 2008). بررسی تنوع ژنتیکی و ارزیابی ساختار جمعیتی برای ارزیابی بهره‌وری و مدیریت بازدارنده‌های ژنتیکی حیوان‌های مزرعه ضروری است. ساختار ژنتیکی بین و درون جمعیت‌ها بازتاب رخدادهای تاریخی گذشته و اخیر، مهاجرت‌ها، اختلاط، نشانه‌های رانش تصادفی و انتخاب طبیعی است (McTavish & Hillis, 2014). در ژنتیک جمعیت، مشخص بودن ساختار جمعیتی، می‌تواند برای ردیابی تاریخچه جمعیت تحت بررسی استفاده شود (Cavalli-Sforza & Feldma, 2003; Lin *et al*., 2010; McTavish & Hillis, 2014). در ژنتیک پزشکی، شناسایی زیر ساختارها و اختصاص دادن افراد به زیرجمعیت‌ها، مرحلهٔ مهم در انجام بررسی‌های ارتباطی برای مشخص کردن پایهٔ ژنتیکی بیماری‌های پیچیده است (Marchini *et al*., 2004; Ziv & Burchard, 2003). گونه‌هایی که پراکندگی جغرافیایی دارند، اغلب تنوع ژنتیکی شایان‌ملاحظه‌ای میان جمعیت‌ها نشان می‌دهند. برآورد زمان و گسترهٔ واگرایی[2] و جریان ژنی میان چنین جمعیت‌هایی برای فهم ساختار ژنتیکی و تفاوت در ژنگان افراد مهم است (Epps *et al*., 2013). به‌طورکلی داده‌های ژنوتیپی با تراکم بالا، امکان تقسیم‌بندی دقیق‌تری از تنوع ژنتیکی گونهٔ به نژادهای متمایز را می‌دهد. در بررسی‌های ارتباطی در سطح ژنگان (GWAS) مشخص کردن تفاوت بین گروه‌های case-control و افرادی که در موقعیت‌های جغرافیایی مختلف هستند یا در آزمایشگاه‌های مختلف تعیین ژنوتیپ شده‌اند امری ضروری است (Liu *et al*., 2013; Price *et al*., 2010). اگر این تفاوت‌ها کم هستند، این گروه‌ها می‌توانند برای به دست آوردن توانمندی بیشتری ادغام شوند (Liu *et al*., 2013).

یادگیری ماشین[3]

روش‌های شناسایی ساختار جمعیت هرکدام وابسته به فرضیه‌هایی بوده و محدودیت‌هایی دارند، مانند روش‌های مبتنی بر مدل ازجمله Structure و Admixture که نیازمند توزیع پیشین برای فراسنجه‌های مدل و متکی به اطلاعات انساب هستند و به‌طورمعمول فرض تعادل هاردی واینبرگ را برای جامعه مدنظر می‌گیرند (Alexander *et al*., 2009; Gao & Starmer, 2007). همچنین روش‌های مبتنی بر مدل محاسبه‌های بسیار گسترده‌ای دارند و در عمل کاربرد آن برای شمار زیادی نشانگر غیرممکن می‌شود (Ma & Amos, 2010). در بررسی ساختار جمعیتی در جوامع با اختلاط جمعیتی زیاد با روش‌های تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA)[4]

1. Single nucleotide polymorphism

۲. واگرایی اشاره به جدا شدن دو نژاد از هم دارد.

3. Machine learning

4. Principal Component Analysis

و STRUCTURE، بیان شد که هر دو روش برای کاربرد روی داده‌های بزرگ مناسب نیستند (Limpiti *et al.*, 2011). روش‌های با نظارت، ابزار مناسب‌تری برای تعیین تفاوت بین جمعیت‌ها هستند. این روش‌ها برای دسته‌بندی جمعیت‌ها استفاده می‌شوند و می‌توانند تفاوت بین دو جمعیت که در مناطق مختلف هستند را به‌خوبی تشخیص دهند درحالی‌که روشی مانند PCA چنین توانایی را ندارد (Bridges *et al.*, 2011). همچنین پیشنهاد شده است که این روش هنگامی‌که افراد به جمعیت‌های از پیش تعریف‌شده دسته‌بندی می‌شوند، در کنترل کیفیت برای بررسی‌های GWAS[1]، روش بهتری باشد (Bridges *et al.*, 2011). یادگیری ماشینی یکی از شاخه‌های گسترده و پرکاربرد هوش مصنوعی است. روش‌های یادگیری ماشینی با استفاده از اطلاعات مجموعهٔ آموزشی یا یادگیری، پیش‌بینی‌های آینده را بر پایهٔ الگو یا قواعد یاد گرفته‌شده انجام می‌دهند. مسئلهٔ دسته‌بندی[2] یکی از مسائل اصلی مطرح‌شده در یادگیری ماشینی است و بسیاری از مسائل را می‌توان به‌صورت یک مسئلهٔ دسته‌بندی مطرح کرده و حل کرد (Larrañaga *et al.*, 2006; Vapnik, 1998). از سوی دیگر در یادگیری ماشینی نیز روش‌های مختلفی برای حل مسئلهٔ دسته‌بندی صورت گرفته است. یکی از روش‌هایی که هم‌اکنون به‌صورت گسترده برای مسئلهٔ دسته‌بندی استفاده می‌شود، روش ماشین بردار پشتیبان (SVM)[3] است (Guinand *et al.*, 2002). به‌طورکلی روش‌های یادگیری ماشینی برای مشکل دسته‌بندی ژنتیکی استفاده می‌شوند. چنین روش‌هایی از الگوریتم‌های خودکار برای تقلید قابلیت‌های یادگیری از مغز حیوانات استفاده می‌کنند. آن‌ها در تجزیهٔ داده‌های پیچیده در بسیاری از رشته‌های علمی بسیار سودمند بوده‌اند (Bridges *et al.*, 2011).

الگوریتم SVM

الگوریتم SVM یا ماشین بردار پشتیبانی یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند (Cortes & Vapnik, 1995). این روش بر پایهٔ این فرض عمل می‌کند که هیچ‌گونه اطلاعی از چگونگی توزیع مجموع داده‌ها وجود ندارد (Steinwart & Christmann 2008). هدف این الگوریتم تشخیص و متمایز کردن الگوهای پیچیده در داده‌ها و دسته‌بندی آن‌ها است (Hastie *et al.*, 2005). درصورتی‌که داده‌ها به‌صورت خطی (طوری‌که با یک یا چند خط بتوان فاصله‌ای بین داده‌ها ایجاد کرد) جداپذیر باشند، ابر صفحه‌ای با بیشترین حاشیه را به دست می‌آورد تا دسته‌ها را جدا کنند و در مواردی که داده‌ها به‌صورت خطی جداپذیر نباشند داده‌ها را به فضای با ابعاد بیشتر نگاشت می‌کند تا در فضای جدید به‌صورت خطی از هم جداسازی شود (Vapnik & Vapnik, 1998). از ویژگی‌های مهم SVM طبقه‌بندی داده‌ها بر پایهٔ کمینه‌سازی خطای ساختاری یا همان خطای آزمایش است (Hsu *et al.*, 2003). اغلب طبقه‌بندی‌های دیگر بر پایهٔ کمینه‌سازی خطای تجربی یا همان خطای آموزش عمل می‌کنند (Boser *et al.*, 1992; Hsu *et al.*, 2003). در این پژوهش برای بررسی ساختار جمعیت گاومیش‌های بومجور آذری و شمالی، روش SVM اجرا شد. ما انتظار داریم این روش به علت استفاده از اطلاعات اولیه، جداسازی بهتری با درستی بالا نشان دهد. هدف این پژوهش دانستن این است که آیا افراد مناطق مختلف درون بومجورها و افراد دو بومجور مناطق مختلف بر پایهٔ ساختار ژنتیکی‌شان قابل جداسازی هستند و روش SVM با چه درستی می‌تواند جمعیت‌ها را از هم جداسازی کند.

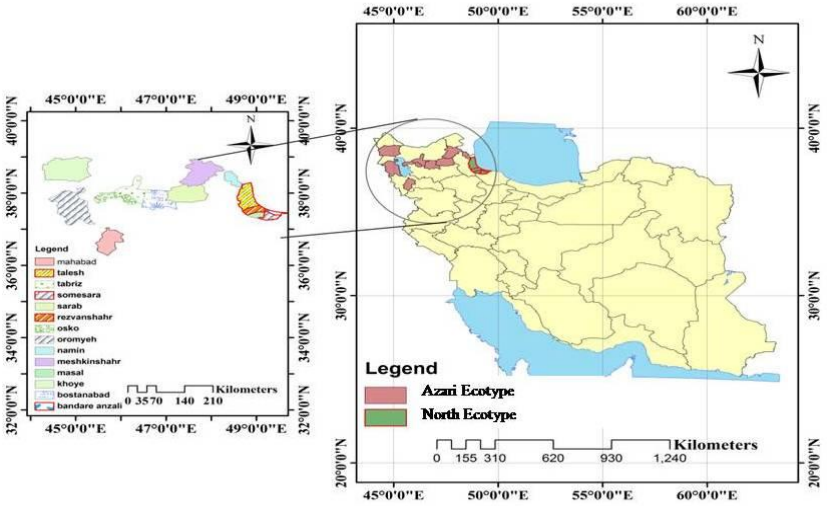
مواد و روش‌ها

نمونه‌های حیوانی و تعیین ژنوتیپ

نمونه‌ها از گله‌هایی که تحت نظام ثبت شجره و رکوردگیری مرکز اصلاح‌نژاد قرارگرفته بودند گردآوری شد. عامل‌هایی که در انتخاب حیوانات موردتوجه بود ازجمله انتخاب حیوانات غیر خویشاوند و انتخاب حیوانات از مناطق مختلف استان‌ها بود. نمونه‌برداری از استان‌های آذربایجان غربی (از سه شهر)، آذربایجان شرقی (از پنج شهر)، اردبیل (از دو شهر) و گیلان (از

1. Genome Wide Association Study
2. Classification
3. Support Vector Machine

هفت شهر) انجام گرفت. در کل ۲۶۲ نمونه به ترتیب ۶۸، ۶۵، ۵۴ و ۷۳ نمونه از استان‌های آذربایجان غربی، اردبیل، آذربایجان شرقی و گیلان گردآوری شد (شکل۱). استخراج DNA ژنگانی از ریشهٔ مو و خون با روش بهینهٔ نمکی انجام شد. نمونه‌ها برای انجام مراحل بعدی توالی‌یابی به آزمایشگاه ژنگانی (ژنومیک) مرکز تحقیقات پادانو (Parco Tecnologico Padano) کشور ایتالیا منتقل شدند. نمونه‌ها با استفاده از تراشه‌های Array Axiom® Buffalo Genotyping 90K مربوط به شرکت افی‌متریکس تعیین ژنوتیپ شدند. این آرایه‌ها امکان تعین ژنوتیپ بیش از ۸۵ هزار جایگاه نشانگری SNP را فراهم می‌آورند.



شکل۱. توزیع جغرافیایی نمونه‌گیری‌های انجام‌شده از مناطق مختلف

Figure1.The geographical distribution of samples taken from different areas

مراحل پالایش داده‌های به‌دست‌آمده از تعیین ژنوتیپ برای انجام تجزیه‌های نهایی

برای اطمینان از کیفیت داده‌های ناشی از تعیین ژنوتیپ، در تجزیه‌های نهایی مراحل مختلف پالایش (فیلتراسیون) روی داده‌های اولیه با استفاده از نرم‌افزار Plink، اعمال شد و حیوانات با بیش از ۵ درصد ژنوتیپ ازدست‌رفته از تجزیه‌های بعدی کنار گذاشته شدند. به دلیل اینکه نمونه‌هایی که کیفیت بالایی ندارند، احتمال بیشتری دارد که با داده‌های گم‌شده همراه باشند و خطای ژنوتیپ در آن‌ها بالا رود (Barendse *et al.*, 2009). دو عامل کمترین فراوانی همردیف ژنی (MAF)[1] و درصدی از نمونه‌ها که برای آن نشانگر تعیین ژنوتیپ شده‌اند (Call rate) برای هر SNP محاسبه وSNPها برای حیوانات دارای Call rate و MAF به ترتیب کمتر از ۹۵ درصد و ۱ درصد حذف شدند. درنهایت برای SNPهای باقی‌مانده، آن‌هایی که تعادل هاردی-واینبرگ نداشتند، به‌عنوان معیاری از خطای تعیین ژنوتیپ (Teo *et al.*, 2007) کنار گذاشته شدند. برای به دست آوردن سطح معنی‌داری در این آزمون از تصحیح بنفرونی ($\beta=\alpha/n$) استفاده شد. بنابراین پس از تعیین ژنوتیپ، عمل غربالگری روی SNPها انجام وSNPهای دارای اطلاعات وارد مرحلهٔ تجزیه‌وتحلیل شدند. در این بررسی کنترل کیفیت اولیه روی داده‌ها توسط شرکت پادانو انجام شد که پس از کنترل کیفیت اولیه، چهار نمونه در جریان تعیین ژنوتیپ (دو نمونه از استان اردبیل و دو نمونه از استان گیلان) با بیش از ۵ درصد ژنوتیپ گم‌شده حذف شدند. ۸۸۵۵ SNP به‌دلیل MAF کمتر از ۱ درصد، ۳۳۶ SNP به دلیل نبود تعادل هاردی-واینبرگ در سطح ۵ درصد و ۱۹ SNP به خاطر موقعیت ناشناخته حذف شدند. درنهایت ۲۵۸ حیوان با ۶۴۷۵۰ SNP، مراحل کنترل کیفیت را با MAF>0.01 و call rate>0.99 گذراندند و همهٔ

1. Minor Allele Frequency

SNPهای باقی‌مانده در سطح ۵درصد در تعادل هاردی- واینبرگ بودند.

محاسبهٔ Fst با استفاده از روش ویر و کوکرهام و ترسیم Neighbor-joining

شاخص ثابت Fst، تفاوت ژنتیکی بین زیرجمعیت‌ها را بر پایهٔ نشانگرهای چند شکل اندازه‌گیری می‌کند که نخستین بار توسط رایت توصیف شد (Wright, 1949). و نقشی مهمی در بررسی‌های ژنتیکی تکاملی دارد و به‌صورت زیر محاسبه می‌شود.

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

که H_T ناخالصی (هتروزیگوسیتی) مورد انتظار برای کل جمعیت است و چنین محاسبه می‌شود.

$$H_T = ۱ - \sum(\overline{p^۲} - \overline{q^۲})$$

و H_S ناخالصی مورد انتظار در زیرجمعیت‌ها است، که چنین محاسبه می‌شود.

$$H_S = \frac{\sum_{i=۱}^{n} H_{expi} \times n_i}{N_{Total}}$$

که $\overline{p}$ و $\overline{q}$ فراوانی همردیف‌های ژنی A1 و A2 کل جمعیت و ni اندازهٔ نمونهٔ زیرجمعیت‌های i است. دامنهٔ Fst از صفر (بدون تفاوت) تا یک (تفاوت کامل، که جمعیت‌ها برای همردیف‌های ژنی مختلف ثابت می‌شوند) است. در عمل مقادیر ۰ تا ۰/۰۵ برای F_{ST} نشان‌دهندهٔ تمایز کم، مقادیر ۰/۰۵ تا ۰/۱۵ نشان‌دهندهٔ تمایز متوسط و بیشتر از ۰/۱۵ تمایز بالا را نشان می‌دهد (Wright, 1969).

روش درخت اتصال همسایگی (NJ)[1] روش خوشه‌بندی بر پایهٔ فاصلهٔ ژنتیکی است که توسط Saitou & Nei (1987) ارائه شد. الگوریتم NJ با یک ساختار شبیه تنهٔ درخت آغاز می‌شود و مجاورین درست (شاخه‌هایی که ازلحاظ فاصله نزدیک‌ترند) را با کمینه کردن مجموع طول همهٔ شاخه‌ها می‌یابد. از این الگوریتم برای رسم درختوارهٔ تبارزایی (فیلوژنیک) بر مبنای فاصله‌های بین افراد توسط نرم‌افزار R استفاده شد.

1. Neighbor-joining

تجزیه‌وتحلیل‌های آماری

آماده‌سازی داده‌ها

برای گسترش یک مدل ماشین بردار پشتیبان، داده‌ها به دو دستهٔ آموزش و آزمایش تقسیم‌بندی می‌شوند. مدل موردنظر توسط داده‌های مجموعهٔ آموزش، آموزش داده می‌شود و کارایی مدل در پیش‌بینی کمیت موردنظر به کمک داده‌هایی که در طول آموزش مدل توسط مدل تجربه نشده‌اند (مجموعه داده‌های آزمایش)، بررسی می‌شود (Kohavi, 1995). یک روش آماده‌سازی مجموعهٔ داده‌ها، تقسیم کل نمونه‌ها به دو دستهٔ آموزش و آزمایش به‌صورت تصادفی است. راهکار دیگری که برای آماده‌سازی داده‌ها وجود دارد استفاده از روش اعتبارسنجی k باره است. در این روش کل داده‌ها به K دستهٔ نزدیک به هم‌اندازه تقسیم می‌شوند. (K-1) دسته برای آموزش مدل و دستهٔ باقی‌مانده برای آزمایش مدل به کار می‌رود. به‌این‌ترتیب، به شمار K مرتبهٔ مدل آموزش و آزمایش می‌شود. سودمندی این روش این است که سرانجام همهٔ نمونه‌های موجود در هر دو فرآیند آموزش و آزمایش شرکت می‌کنند.

انتخاب فراسنجه

روش SVM با پکیج e1071 اجرا می‌شود که مقادیر ϒ و C برای گسترش مدل SVM با تابع () tune که ده بار اعتبارسنجی انجام می‌دهد، تنظیم می‌شود. انواع توابع کرنل در این تحقیق اجرا شد. در مرحلهٔ اول افراد سه استان از بوم‌جور آذری با افراد منطقهٔ بوم‌جور گیلانی به‌عنوان چهار کلاس در نظر گرفته‌شده و تجزیه می‌شوند و در مرحلهٔ بعد استان‌های بوم‌جور آذری باهم ادغام می‌شوند و با بوم‌جور گیلانی به‌عنوان دو کلاس در نظر گرفته‌شده و تجزیه می‌شوند.

ارزیابی اعتبار مدل

برای ارزیابی روش موردنظر در این تحقیق، از معیارهای ارزیابی همچون ماتریس اختلاط و منحنی راک برای اطمینان از درستی روش SVM استفاده می‌شود (Stehman, 1997; Swets, 1988). در مورد گروه‌بندی با توجه به شمار افرادی که در گروه درستی

قرار می‌گیرند درستی به دست می‌آید و هرچقدر این عدد بزرگ‌تر باشد درستی بیشتر و گروه‌بندی درستی صورت گرفته است. به عبارتی افرادی که در گروه‌های نادرست قرار می‌گیرند به‌عنوان خطای گروه‌بندی در نظر گرفته می‌شوند و میزان خطای گروه‌بندی[1] معیاری برای ارزیابی اعتبار مدل است.

ماتریس اختلاط

ارزیابی نتایج طبقه‌بندی یکی از مراحل مهم پس از هر فرآیند طبقه‌بندی است. یکی از روش‌های متداول ارزیابی طبقه‌بندی، استفاده از مجموعهٔ نمونه‌های آزمون و تشکیل ماتریس خطا است (جدول۱). این ماتریس چگونگی عملکرد الگوریتم دسته‌بندی را با توجه به مجموعهٔ دادهٔ ورودی به جداسازی انواع دسته‌های مسئله دسته‌بندی، نمایش می‌دهد. منفی اشتباه (FN)، شمار رکوردهایی است که دستهٔ واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی دستهٔ آن‌ها را به‌اشتباه منفی تشخیص داده است. منفی صحیح (TN)، شمار رکوردهایی است که دستهٔ واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دستهٔ آن‌ها را به‌درستی منفی تشخیص داده است. مثبت صحیح (TP)، شمار رکوردهایی است که دستهٔ واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دستهٔ آن‌ها را به‌درستی مثبت تشخیص داده است. منفی اشتباه (FP)، شمار رکوردهایی است که دستهٔ واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی دستهٔ آن‌ها را به‌اشتباه مثبت تشخیص داده است.

جدول ۱. ماتریس اختلاط

Table1. Confusion Matrix

	Predicted Positive	Predicted Negative
True Negative	FN[2]	TN[3]
True Positive	TP[4]	FP[5]

1. Classification error rate
2. False Negative
3. True Negative
4. True Positive
5. False Positive

مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی دقت یا میزان دسته‌بندی[6] است که این معیار دقت کل یک دسته‌بند را محاسبه می‌کند. درواقع این معیار مشهورترین و عمومی‌ترین معیار محاسبهٔ کارایی الگوریتم‌های دسته‌بندی است که نشان می‌دهد، دسته‌بند طراحی‌شده چند درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی دسته‌بندی کرده است. اطلاعاتی که از این جدول به‌دست می‌آید شامل موارد زیر است:

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (۱)$$

$$\text{Fall}-\text{out} = \text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}} \quad (۲)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}} \quad (۳)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (۴)$$

دقت، نسبت کل شمار پیش‌بینی‌های درست را نشان می‌دهد. حساسیت یا Sensivity نسبت موارد مثبتی است که به‌درستی شناسایی شده است و ویژگی یا Specificity نسبت موارد منفی است که به‌طور درست دسته‌بندی‌شده‌اند. نرخ مثبت اشتباه یا FPR، نسبت موارد منفی است که به‌طور نادرستی دسته‌بندی شده است.

منحنی ROC[7]

یکی از روش‌های مناسب برای ارزیابی نتایج به‌دست‌آمده از یک طبقه‌بندی‌کننده و ارزیابی میزان قابلیت آن در شناسایی طبقهٔ موردنظر استفاده از منحنی عملیاتی دریافت‌کننده به‌منظور بررسی حساسیت روش است. منظور از حساسیت، ارتباط بین میزان یاخته‌های درست طبقه‌بندی‌شده و موارد نادرست است. هرچه میزان انحراف از خط مبنا برای یک بیشتر باشد، کارایی طبقه‌بندی ROC در شناسایی آن طبقه بیشتر است. افزون بر بررسی روند نمودار طبقهٔ موردنظر، سطح زیر آن نمودار نیز محاسبه می‌شود. این گستره بیانگر احتمال این است که یک

6. Classification Accuracy – Rate
7. Receiver Operating Characteristic

یاختهٔ انتخاب‌شده به‌طور تصادفی، صحیح طبقه‌بندی می‌شود و هرچه بیشتر باشد، قابلیت اطمینان روش یادشده را نشان می‌دهد (Hand, 2009; Swets, 1988). این منحنی درواقع تغییرپذیری TPR در برابر FPR است. برای رسم این منحنی از پکیج pROCR استفاده شد.

نتایج و بحث

نتایج کنترل کیفیت

در این بررسی کنترل کیفیت اولیه روی داده‌ها توسط شرکت پادانو انجام گرفته بود که اطلاعات مربوط به کنترل کیفیت این شرکت در قسمت اطلاعات تکمیلی ارائه شده است. کنترل کیفیت را روی SNP۶۴۷۵۰ به‌دست‌آمده از کنترل کیفیت اولیه اجرا شد که در آغاز ۱۹ SNP به دلیل موقعیت ناشناخته حذف شدند و در مراحل مختلف کنترل کیفیت روی SNPهای باقی‌ماندهٔ ۷ SNP با MAF کمتر از ۱ درصد حذف شدند و ۵ SNP هم به دلیل انحراف از تعادل هاردی-واینبرگ از تجزیه‌های نهایی حذف شدند و درمجموع ۲۵۸ حیوان از چهار استان مختلف از دو بومجور با ۶۴۷۱۹ SNP وارد مرحلهٔ تجزیهٔ نهایی شدند.

تجزیهٔ آماری

تفاوت ژنتیکی میان نواحی جغرافیایی

برای ارزیابی تفاوت ژنتیکی میان نواحی مختلف دو بومجور، درخت اتصال همسایه (NJ) رسم شد (شکل‌های ۲ و ۳) و Fst (جدول۲) بین هر جفت از زیر جمعیت‌ها محاسبه شد.

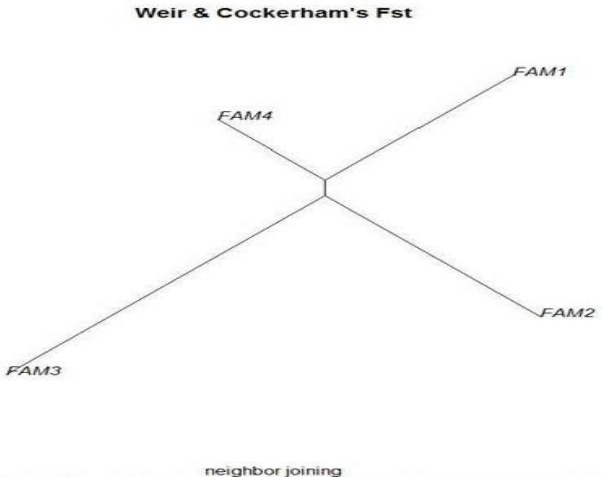
جدول۲. فاصلهٔ Fst میان جمعیت‌های چهار استان از دو بومجور

Table 2. Fst distance between populations of 4 provinces from the two ecotypes

	FAM2 (Guilan)	FAM3 (Ardabil)	FAM4 (East Azarbaijan)
FAM1	0.008317581	0.009894526	0.005687029
FAM2		0.010025664	0.006982461
FAM3			0.008061572

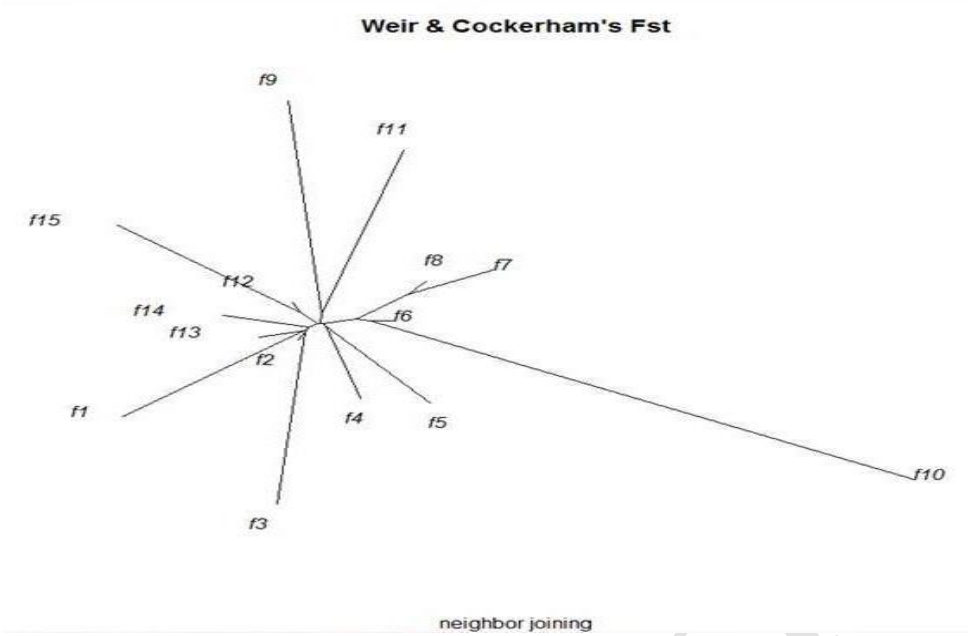
نتایج به‌دست‌آمده نشان می‌دهد که حیوانات استان‌های مختلف تفاوت ژنتیکی کمی داشتند و دراین‌بین تفاوت ژنتیکی گاومیش‌های استان اردبیل با گاومیش‌های استان گیلان نسبت به استان‌های دیگر بیشتر و Fst = ۰/۰۱ به دست آمد.

درختوارهٔ اتصال همسایگی یا پیوند همجواری که برحسب فاصلهٔ ژنتیکی شاخص ثابت Fst رسم شد فاصلهٔ گاومیش‌های چهار استان و شهرهای چهار استان را نشان می‌دهد که جمعیت این استان‌ها ازلحاظ ژنتیکی به هم نزدیک بوده و تفاوت ژنتیکی کمی دارند.



شکل۲. درختوارهٔ ژنتیکی Neighbor-Joining (NJ) برحسب فاصلهٔ ژنتیکی (Fst) جمعیت‌های استان‌ها (FAM1=West Azarbaijan, FAM2= Guilan, FAM3=Ardabil & FAM4=East Azarbaijan)

Figure 2. Neighbor-Joining (NJ) based on F_{ST} of different provinces



شکل۳. درختوارهٔ ژنتیکی Neighbor-Joining (NJ) برحسب فاصلهٔ ژنتیکی (Fst) شهرهای مختلف استان‌ها (f1، f2،f3، f13 و f14 مربوط به شهرهای استان گیلان، f4 و f5 مربوط به شهرهای استان اردبیل، f9، f11،f12 و f15 مربوط به شهرهای استان آذربایجان غربی و f6،f7،f8 و f10 مربوط به شهرستان‌های استان آذربایجان شرقی است.

Figure 3. Neighbor-Joining(NJ) based on F_{ST} of different cities from provinces(f1,f2,f3,f13&f14 cites of guilan, f4&f5 cities of Ardabil, f9,f11,f11,f12&f15 cites of West Azarbayjan, f6,f7,f8&f10 cites of East Azarbayjan).

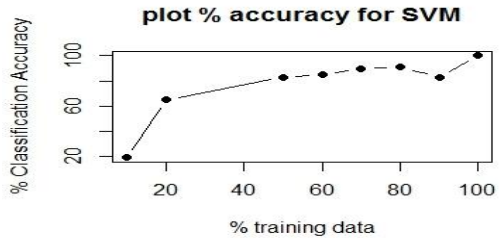
درختوارهٔ اتصال شهرستان‌های مختلف از چهار استان در شکل ۳ ارائه شده است که شهرهای مختلف از چهار استان نیز تفاوت ژنتیکی کمی نشان می‌دهند ولی برخلاف خود چهار استان که به‌خوبی از هم می‌توان جداسازی کرد در اینجا این جداسازی به‌راحتی انجام نمی‌شود.

نتایج مدل آماری حاصل از تجزیه با الگوریتم SVM

با توجه به اینکه درروش SVM دو فراسنجهٔ گاما و C بایستی تنظیم شوند، بهترین عملکرد فراسنجه‌های گاما و C با تابع ()tuning در برنامهٔ R با انجام ده بار اعتبارسنجی به دست آمد و انواع کرنل‌های این روش ازجمله کرنل‌های خطی و شعاعی و حلقه‌ای برای مدل SVM بررسی شد که درنهایت درستی به‌دست‌آمده از انواع کرنل‌ها نزدیک به هم بود و تفاوت معنی‌داری نداشتند.

تجزیهٔ چهار استان از دو بوم‌جور

در این مرحله از تجزیهٔ چهار استان از دو بوم‌جور به‌عنوان چهار کلاس در تجزیه وارد شدند و نتایج به‌دست‌آمده با داده‌های آموزشی و آزمایشی با داده‌های آموزشی از ده تا ۱۰۰ به‌صورت نمودار زیر است که بهترین مقدار دادهٔ آموزشی (۸۰:۲۰) با درستی نزدیک ۹۱ درصد بود (شکل ۴) که با این درستی افراد مناطق مختلف قابل جداسازی هستند. نتایج ناشی از اعتبارسنجی با k=10 نیز درستی ۹۲ نشان داد.



شکل ۴. نتایج درستی گروه‌بندی برای نسبت‌های مختلف داده‌های آموزشی و آزمایشی

Figure 4. The results of classification accuracy for different ratios of training and test data

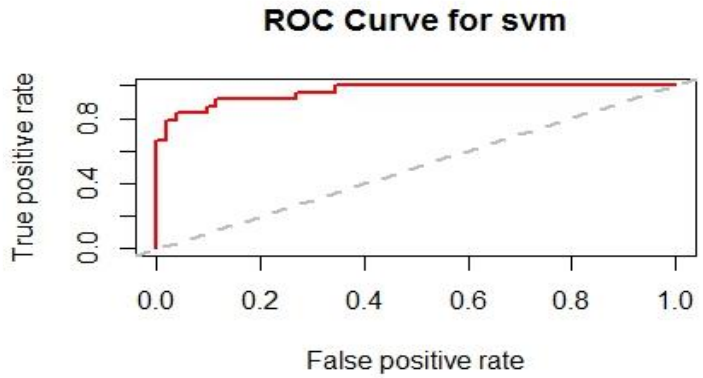
منحنی راک نیز که معیاری دیگر برای ارزیابی مدل موردنظر است سطح زیر منحنی، درستی کلاسه‌بندی‌کننده را برای دسته‌بندی درست تا ۹۸

درصد تأیید کرد (شکل ۵). افزون بر منحنی راک، شاخص‌های دقت و درستی نیز محاسبه شدند که نتایج محاسبهٔ این شاخص‌ها و مؤلفه‌های مربوط به آن‌ها در جدول ۳ ارائه شده است.

جدول ۳. خلاصهٔ نتایج به‌دست‌آمده از محاسبهٔ شاخص‌های منحنی ROC، درستی کلی و ضریب کاپا در طبقه‌بندی‌کنندهٔ SVM

Table 3. Summary results of the index curve ROC, overall accuracy and kappa coefficient in SVM classifier

Index	F1 (West Azarbaijan)	F2 (Guilan)	F3 (Ardabil)	F4 (East Azarbaijan)
Sensitivity	1.0000	1.0000	0.7143	1.0000
Specificity	1.0000	0.9565	1.0000	0.9375
PosPred Value	1.0000	0.7500	1.0000	0.9091
Neg Pred Value	1.0000	1.0000	0.9048	1.0000
Balanced Accuracy	1.0000	0.9783	0.8571	0.9688
Accuracy		0.9231		
Kappa		0.8923		
AUC		0.9762		



شکل۵. منحنی راک طبقه‌بندی‌کنندهٔ SVM برای چهار استان مختلف

Figure 5. ROC curve of SVM classifier for 4 different provinces

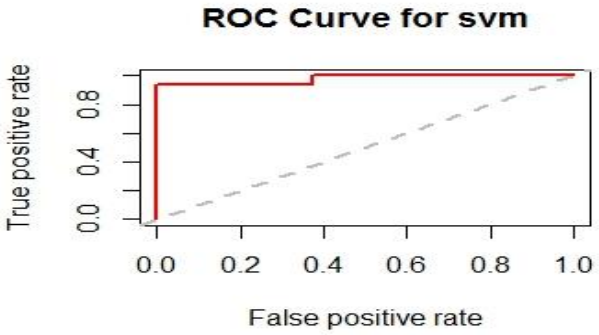
هرکدام از موارد یادشده در جدول که از ماتریس اختلاط به دست می‌آیند به‌نوعی دقت روش موردنظر را نشان می‌دهند. در شکل ۵ نیز با توجه به سیر صعودی منحنی می‌توان از عملکرد بالای سامانهٔ تشخیصی با مجموعه دادهٔ آزمون و حتی هر مجموعه دادهٔ جدید دیگری اطمینان به دست آورد. این روش با توجه به پیش‌بینی‌هایی که انجام می‌دهد، احتمال اختصاص یافتن هر فرد به کلاس مربوط به خود را با درستی ۹۲ درصد می‌دهد.

تجزیهٔ دو بوم‌جور به‌عنوان دو کلاس جداگانه

در مرحلهٔ بعدی، داده‌های سه منطقه از بوم‌جور آذری، یک کلاس در نظر گرفته شد و با بوم‌جور گیلانی به‌عنوان کلاس دیگر تجزیه شد. نتایج به‌دست‌آمده از تجزیهٔ مدل SVM با داده‌های آموزشی ۸۰ به ۲۰ نشان‌دهندهٔ درستی ۹۴درصد بود و نتایج اعتبارسنجی با K=10 درستی ۹۶ درصد نشان داد که گویای این موضوع است که هنگامی افراد مناطق نزدیک به هم در یک کلاس در نظر گرفته می‌شوند جداسازی آسان‌تر و با درستی بالایی انجام می‌شود. منحنی راک نیز تأییدی بر این نتیجه‌گیری بود که سطح زیر منحنی برابر با ۰/۹۸ است (شکل ۶).

به‌دست آوردن ساختار جمعیتی دقیق و جداسازی افراد و زیرجمعیت‌ها با درستی بالا با روش‌های توانمندی مانند روش‌های یادگیری ماشینی، می‌تواند در شرایط گوناگون مانند بررسی‌های ارتباطی و تکاملی و تعیین بازدارنده‌های ژنتیکی سودمند واقع شود. در این تحقیق روش SVM توانست با استفاده از داده‌های ژنوتیپی جداسازی خوبی از زیرجمعیت‌ها و دو بوم‌جور مختلف را داشته باشد، طوری‌که فردی با ژنوتیپ مشخص را با استفاده از این روش می‌توان تعیین کرد که به کدام بوم‌جور و منطقه تعلق دارد. همچنین درصد اختلاط هر فرد به کلاس‌های مربوطه را با پیش‌بینی‌هایی که انجام می‌دهد با درستی ۹۶ درصد، می‌دهد. سطح زیر منحنی

راک AUC به‌عنوان معیار ارزیابی عملکرد روش مورداستفادهٔ عدد بالا و ۹۸ درصد را نشان داد که گویای بالا بودن میزان قابلیت شناسایی گروه‌های موردنظر توسط روش SVM است (جدول ۴).



شکل ۶. منحنی راک طبقه‌بندی‌کنندهٔ SVM برای دو بوم‌جور
Figure6. ROC curve of SVM classifier for 2 ecotypes

جدول ۴. خلاصهٔ نتایج به‌دست‌آمده از محاسبهٔ شاخص‌های منحنی ROC، درستی کلی و ضریب کاپا در طبقه‌بندی‌کنندهٔ SVM

Table 4. Summary results of the index curve ROC, overall accuracy and kappa coefficient in SVM classifier

Index	F1 (Azari Ecotype)
Sensitivity	0.9545
Specificity	1.0000
PosPred Value	1.0000
NegPred Value	0.8000
Balanced Accuracy	0.9773
Accuracy	0.9615
Kappa	0.866
AUC	0.98

نتایج به‌دست‌آمده از Fst به‌عنوان شاخصی از تفاوت ژنتیکی، گویای تفاوت ژنتیکی کم بین جمعیت بوم‌جورها بود و روش SVM با توجه به اطلاعات پیشین از افراد، احتمال‌های مربوط به هر فرد را که با چه درصد درستی ممکن است به کلاس مربوط به خود تعلق گیرند را می‌دهد. الگوریتم SVM برای تجزیهٔ جمعیت‌های زیست‌شناختی (بیولوژیکی) با گروه‌بندی با نظارت برای انتساب افراد که توسط داده‌های توالی مشخص شدند، استفاده شده است. از سوی دیگر با توجه به اینکه روش‌های با نظارت توانایی تشخیص الگوها و انتساب افراد به جمعیت موردنظر را دارند (Brown *et al.*, 2000)، می‌توان گفت که روش SVM با احتمال‌هایی که برای گروه‌بندی هر فرد می‌دهد، قابلیتی را برای کنترل کیفیت داده‌ها با حذف افرادی که با احتمال پایین گروه‌بندی می‌شوند، فراهم می‌آورد. در بررسی حساسیت روش با نظارت (شبکهٔ عصبی و SVM) در مقایسهٔ دسته‌بندی ژنتیکی با استفاده از روش‌های بدون نظارت مانند PCA روی جمعیت‌های مختلف (دو جمعیت اسکاتلندی و یک جمعیت بلغاری)، به‌طور شایان‌توجهی بیشتر از PCA بود (Bridges *et al.*, 2011). همچنین در بررسی روی بیست رقم انگور درستی به‌دست‌آمده از مدل‌های SVM و ANN برای گروه‌بندی این رقم‌ها حدود ۸۷/۲۵ درصد بود که نشان‌دهندهٔ قابلیت اطمینان بالای این مدل‌های گروه‌بندی است (Gutiérrez *et al.*, 2015). در تجزیهٔ تصویر سه‌بعدی برای دسته‌بندی بیماری‌های استخوانی، ویژگی‌های ریخت‌شناختی (مورفولوژیکی)، ساختاری (توپولوژیکی) و مکانیکی استخوان از روش‌های هوش مصنوعی استفاده شد که سامانهٔ استنتاج نروفازی (ANFIS)، ماشین بردار پشتیبان (SVM) و الگوریتم ژنتیک استفاده شدند (Akgundogdu *et al.*, 2010). روش SVM برای گروه‌بندی داده‌های بیان ژنی اعمال شده است و درستی کلی به‌دست‌آمده ۷۸ درصد گزارش شد (Brown *et al.*, 2000). با توجه به این موارد، این روش می‌تواند در بررسی‌های GWAS برای کلاسه‌بندی جمعیت‌ها و کنترل کیفیت و در کاربردهای ژنتیکی ازجمله تعیین جمعیت‌ها و زیرجمعیت‌ها به‌ویژه در ارزیابی‌های ژنتیکی و ارزیابی‌های ژنگانی بسیار اهمیت دارد.

نتیجه‌گیری کلی

با روش SVM، افراد با ژنوتیپ مشخص، می‌توانند با درستی بالایی به نژاد، منطقه یا گله‌ای که به آن تعلق دارند، اختصاص یابند به‌طوری‌که اگر شماری نمونه با هویت مجهول داشته باشیم می‌توانیم با این روش به نژاد یا جمعیتی که به آن تعلق دارند اختصاص دهیم. با توجه به این امر، اجرای روش‌های دیگری از یادگیری ماشینی و انجام مقایسه‌ها و درستی به‌دست‌آمده از این روش‌ها با در نظر گرفتن زمان محاسبهٔ برنامه، پیشنهاد می‌شود.

سپاسگزاری

از مسئولان مرکز اصلاح نژاد و بهبود تولیدهای دامی کشور و شرکت دانشگاهی و دانش‌بنیان نواندیش البرز که حمایت مالی این پروژه را بر عهده داشتند، تشکر و قدردانی می‌گردد.

REFERENCES

1. Akgundogdu, A., Jennane, R., Aufort, G., Benhamou, C.L. & Ucan, O.N. (2010). 3D image analysis and artificial intelligence for bone disease classification. *Journal of Medical Systems*, 34(5), 815-828.
2. Alexander, D. H., Novembre, J. & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9), 1655-1664.
3. Anonymous.Statical center of Iran (2012). Available from: http: www.amar.org ir/
4. Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). *A training algorithm for optimal margin classifiers*. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory.
5. Bridges, M., Heron, E. A., O'Dushlaine, C., Segurado, R., Morris, D., Corvin, A., ... Consortium, I. S. (2011a). Genetic classification of populations using supervised learning. *PLoS ONE*, *6*(5), e14802.
6. Bridges, M., Heron, E. A., O'Dushlaine, C., Segurado, R., Morris, D., Corvin, A., ... Consortium, I.S. (2011b). Genetic classification of populations using supervised learning.
7. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1), 262-267.
8. Buturovic, L., Cohen, S., He, Z., Eggenberger, M., Nacci, D. & Petkovic, D. Supervised Classification of Genetic Sequences for Population Analysis.
9. Cavalli-Sforza, L. L. & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *nature genetics*, 33, 266-275.
10. Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
11. Epps, C. W., Castillo, J. A., Schmidt-Küntzel, A., du Preez, P., Stuart-Hill, G., Jago, M. & Naidoo, R. (2013). Contrasting historical and recent gene flow among African buffalo herds in the Caprivi Strip of Namibia. *Journal of Heredity*, ess142.
12. Fernández, M. E., Goszczynski, D. E., Lirón, J. P., Villegas-Castagnasso, E. E., Carino, M. H., Ripoli, M. V., ... Giovambattista, G. (2013). Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genetics and molecular biology*, 36(2), 185-191.
13. Gao, X. & Starmer, J. (2007). Human population structure detection via multilocus genotype clustering. *BMC genetics*, 8(1), 34.
14. Guerard, E., Heyer, E. & Manni, F. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Human biology*, 76(2), 173-190.
15. Guinand, B., Topchy, A., Page, K., Burnham-Curtis, M., Punch, W. & Scribner, K. (2002). Comparisons of likelihood and machine learning methods of individual classification. *Journal of Heredity*, 93(4), 260-269.
16. Gutiérrez, S., Tardaguila, J., Fernández-Novales, J., Diago, M. P. & Scali, M. (2015). Support Vector Machine and Artificial Neural Network Models for the Classification of Grapevine Varieties Using a Portable NIR Spectrophotometer. *PLoS ONE*, 10(11), e0143197.
17. Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1), 103-123.
18. Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83-85.
19. Hsu, C.-W., Chang, C.-C. & Lin, C.-J. (2003a). A practical guide to support vector classification.
20. Hsu, C.-W., Chang, C.-C. & Lin, C.-J. (2003b). A practical guide to support vector classification.
21. https://cran.r-project.org/web/packages/e1071/index.html.

22. Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.
23. Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., ... Comas, D. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology,* 18(16), 1241-1248.
24. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... Pérez, A. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics,* 7(1), 86-112.
25. Limpiti, T., Intarapanich, A., Assawamakin, A., Shaw, P. J., Wangkumhang, P., Piriyapongsa, J., ... Tongsima, S. (2011). Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. *BMC bioinformatics,* 12(1), 255.
26. Lin, B. Z., Sasazaki, S. & Mannen, H. (2010). Genetic diversity and structure in Bos taurus and Bos indicus populations analyzed by SNP markers. *Animal science journal,* 81(3), 281-289.
27. Liu, L., Zhang, D., Liu, H. & Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC bioinformatics,* 14(1), 1.
28. Ma, J. & Amos, C. I. (2010). Theoretical formulation of principal components analysis to detect and correct for population stratification. *PLoS ONE,* 5(9), e12510.
29. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature genetics,* 36(5), 512-517.
30. McTavish, E. J. & Hillis, D. M. (2014). A genomic approach for distinguishing between recent and ancient admixture as applied to cattle. *Journal of Heredity, 105*(4), 445-456.
31. Naserian, A. A. & Saremi, B. (2010). Water buffalo industry in Iran. *Italian Journal of Animal Science,* 6(2s), 1404-1405.
32. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics,* 11(7), 459-463.
33. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution,* 4(4), 406-425.
34. Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment,* 62(1), 77-89.
35. Steinwart, I. & Christmann, A. (2008). *Support vector machines*: Springer Science & Business Media.
36. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science,* 240(4857), 1285-1293.
37. Thomas, D. C. & Witte, J. S. (2002). Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiology Biomarkers & Prevention,* 11(6), 505-512.
38. Vapnik, V. N. & Vapnik, V. (1998). *Statistical learning theory* (Vol. 1): Wiley New York.
39. Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution,* 34(3), 275-306.
40. Wacholder, S., Rothman, N. & Caporaso, N. (2002). Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiology Biomarkers & Prevention,* 11(6), 513-520.
41. Wright, S. (1949). The genetical structure of populations. *Annals of eugenics,* 15(1), 323-354.
42. Wright, S. (1969). Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies.
43. Ziv, E. & Burchard, E.G. (2003). Human population structure and genetic association studies. *Pharmacogenomics,* 4(4), 431-441.

Iranian Journal of Animal Science, Vol. 47, No. 2, Summer 2016

Genetic classification of Azari and North ecotype Buffalo population using SVM method

Zahra Azizi[1], Hossein Moradi Shahrbabak[2*], Mohammad Moradi Shahrbabak[3], Seyed Abbas Rafat[4] and Jalil Shodja[5]

1, 4, 5. Ph.D. Student, Associate Professor and Professor, Department of Animal Sciences, Faculty of Agricultural Sciences, University of Tabriz, Iran

2, 3. Assistant Professor and Professor, Department of Animal Sciences, University College of Agriculture & Natural Resources, University of Tehran, Karaj, Iran

(Received: Jan. 5, 2016 - Accepted: Apr. 13, 2016)

ABSTRACT

The purpose of this research was to classify buffaloes from different areas of the two Azari (West and East Azarbayjan and Ardabil provinces) and North (Guilan province) ecotypes using support vector machine method. A total of 258 buffalo were sampled and genotyped using the Axiom Buffalo 90K Genotyping Array at the Parco Technologic Padano lab in Italy. Two metric methods of cross validation and the area under the receiver operating characteristic (AUC) were used to determine the predictive performance of support vector machine (SVM) to classify individuals. The results of cross validation and methods for classifying different regions of the two ecotypes (4 provinces) were 92% and 96%, respectively that showed despite the difficulty of identifying individuals from provinces close to each other, support vector machine (SVM) method shows higher accuracy in assigning animals to their herds. Result of two ecotypes showed accuracy about 96% and 98% which represents the better ability to separate the two ecotypes. Machine learning method provides predictions for classification of each individual which can be efficient in quality control and genetic studies.

Keywords: Buffalo, classification, SNPChip 90K, support vector machine.

* Corresponding author E-mail: hmoradis@ut.ac.ir Tel: +98 26 32248082